Chapter 5

Identifying Peptide-based Antibacterial Drugs effective against ESKAPEE pathogens using Artificial Intelligence

To guarantee survival in the presence of antibiotics, pathogens make several changes in their genome, leading to antimicrobial resistance (AMR). Amongst the drug-resistant bacteria, the ESKAPEE (*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, Enterobacter spp., and Escherichia coli*) pathogens pose a major threat. Antibacterial peptides (ABPs) are a family of natural peptides that play a critical role in the innate immune systems of organisms. Due to the problem of AMR, ABPs have recently garnered considerable interest as a potential alternative to available antibiotics. However, finding ABPs from natural sources that can target all ESKAPEE pathogens at low concentration is time-consuming and costly. As a result, to undertake a preliminary screening of natural sources to discover potential ABPs effective against ESKAPEE pathogens, the

in-silico tool is needed. Thus, this chapter introduces an artificial intelligence-based framework named ESKAPEE-MICpred, which has been made available as a web app at https://eskapee-micpred.anvil.app/. This app provides the minimum inhibitory concentration values for the peptide against ESKAPEE pathogens, which will aid wet lab researchers in the fight against antibacterial resistance by accelerating the discovery of peptide-based antibacterial medications.

5.1 Introduction

To guarantee survival in the presence of antibiotics; pathogens make several changes in their genome, which leads to antimicrobial resistance (AMR). The World Health Organization (WHO) categorizes bacteria into three categories of priority: critical, high, and medium, according to the urgency of the need to develop new antibiotics to combat these pathogens. Amongst the drug-resistant bacteria, the ESKAPEE (*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, Enterobacter spp., and Escherichia coli*) pathogens pose a major threat, which consists of high to critical WHO-priority pathogens [100, 91, 92, 101, 102]. Antibacterial peptides (ABPs) are a family of natural peptides that play a critical role in the innate immune systems of organisms. These ABPs offer several advantages over commonly used antibiotics (naturally produced, destroying pathogens in various ways, and having minimal side effects). The therapeutic uses of ABPs are provided in [86, 87, 88]. Due to the problem of AMR, ABPs have recently garnered considerable interest as a potential alternative to available antibiotics [84]. However, finding ABPs from natural sources is time taking and costly. Thus, wet lab researchers employ various tools available in the public domain to screen promising ABPs rapidly.

The main drawback of existing tools is that they do not provide the minimum inhibitory concentration (MIC) values against the ESKAPEE pathogens for the identified ABP. Therefore, after identifying ABPs, wet lab researchers have to test them against

the ESKAPEE pathogens at different concentrations, leading to a loss of time and money. Apart from this, due to the involvement of time and money, testing all the identified ABPs in the lab at different concentrations is not feasible, which may lead to loss of optimal ABP(s), which can work at low MIC(s) against all the ESKAPEE group of bacteria. To address this, in the current work, we proposed a model named ESKAPEE-MICpred, which provides MIC values for the ABP against ESKAPEE pathogens.

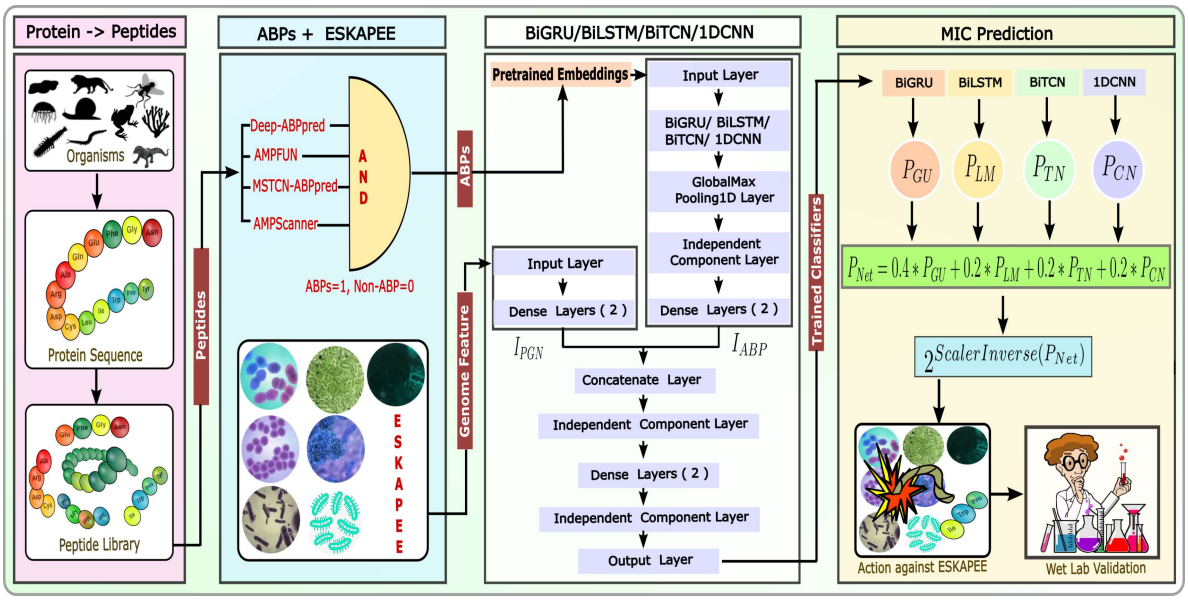

Figure 5.1: Proposed Framework.

Deep learning-based models have demonstrated their efficacy in various challenging applications [103, 104, 105, 106, 107]. In the field of natural language processing and bioinformatics, bidirectional gated recurrent units (BiGRU), bidirectional long short-term memory (BiLSTM), bidirectional temporal convolutional network (BiTCN), and 1-dimensional convolutional neural network (1DCNN) have been utilized in sequence-related tasks. Peptides, which are composed of sequences of amino acids, also benefit from the use of these powerful deep learning models. Moreover, instead of using these algorithms individually, combining them can result in an aggregate prediction that is less noisy and more accurate. As a result, we integrated their results to develop the ESKAPEE-MICpred. To understand the contribution of ensemble learning, we also

performed the ablation studies and found a decrease in the performance on removing the ensemble learning technique. Raw peptides cannot be used to train an algorithm. Therefore, they must be encoded into numerical form before being utilized for training an algorithm. The method utilized to encode raw peptides into numerical form also affects the performance. Therefore, we utilized the concept of transfer learning with the help of pretrained embeddings from seq2vec (PESTV) for encoding the peptides. Authors in [14] learned PESTV by training Embeddings from Language Models (ELMo) on millions of protein sequences from UniRef50.

To understand the contribution of PESTV, we performed the ablation studies, where we replaced the PESTV with different non-pretrained encodings, namely NNAA, PAM250, BLOSUM62, and one hot encoding (OHE). From the ablation studies, we found a decrease in performance on replacing PESTV with other non-pretrained encodings.

For building ESKAPEE-MICpred, we prepared a dataset having 11,266 ABPs using the latest database named DBAASP v3 [64]. The efficacy of ABPs is to be tested against the ESKAPEE pathogens. Therefore seven whole genome sequences (WGS) for ESKAPEE pathogens were also obtained from NCBI [108].

The proposed model has been implemented as a web server to aid wet-lab researchers and is freely available online at https://eskapee-micpred.anvil.app/. As a proof of concept, we identified five ABPs from the antibacterial protein sequences and five ABPs from the therapeutic peptides.The graphical abstract of steps performed for identifying ABPs from the protein sequences are shown in Figure 5.1, and the detailed explanation is provided in Section 5.4. The main contributions of our chapter are summarised as follows:

1. In the current study, we developed a model named ESKAPEE-MICpred to supplement the currently available tools. This model provides MIC values against the ESKAPEE pathogens for the given ABP.

2. For our proposed model, ESKAPEE-MICpred, the Pearson correlation coefficient is greater than 0.8, demonstrating a high positive correlation between the actual and predicted MIC values.
3. As a proof of concept, we identified five ABPs from the antibacterial protein sequences and five ABPs from the therapeutic peptides.
4. To aid the scientific community, the proposed model has been deployed as a web server at https://eskapee-micpred.anvil.app/.

The rest of this chapter is structured as follows. The information regarding the dataset and architecture of the proposed model is presented in Section 5.2. Section 5.3 provides the different experiments conducted and corresponding results obtained. The identification of ABPs from the antibacterial protein sequences and the therapeutic peptides is presented in Section 5.4. The information about the web server is provided in Section 5.5. The conclusion is given in Section 5.6.

5.2 Materials and methods

5.2.1 Dataset

The dataset (D_s) consists of 11,266 ABPs. These ABPs have lengths $\in$ [5,50], MIC values $\in$ (0,1024) µM, are made up of natural amino acids, and each ABP is effective against at least one pathogen of the ESKAPEE group. For creating D_s, we considered the most recent database, DBAASP v3 [64]. It contains ABPs that have experimental MIC values against various bacterial species. The MIC values are available in DBAASP v3 as µM or µg/ml, which were made uniform by converting µg/ml to µM. Due to varying experimental conditions, multiple studies have reported different MIC values for the same ABP. As a result, there are multiple entries for some ABPs, which we reduced to a single entry by taking their mean. Moreover, ABPs were unevenly distributed across a wide range of MIC values 5.3(A), which can affect the training of the model.

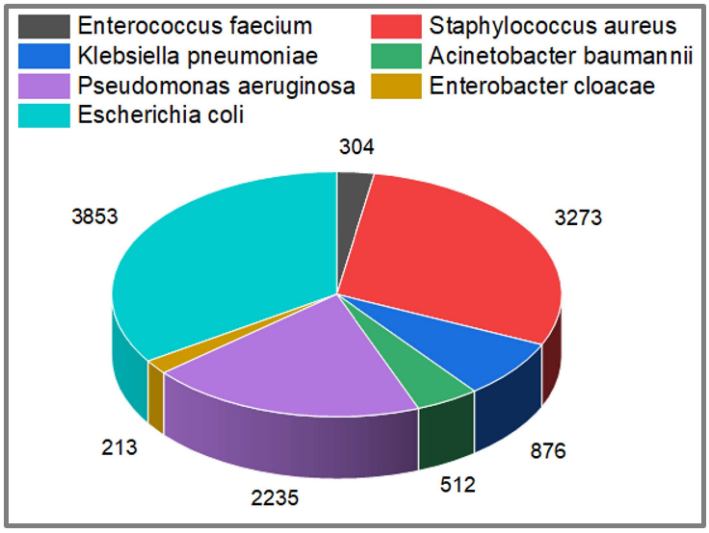

Figure 5.2: Number of peptides active against each bacterial species of the ESKAPEE group.

Therefore, we applied log transformation, which makes ABPs evenly distributed across a narrow range 5.3(B). After log transformation, we performed standardization, which makes the mean=0 and standard deviation=1 5.3(C). We further divided the D_s into three sets, namely Training set (S^{Train}), Validation set (S^{Val}), and Test set (S^{Test}). S^{Train} contains $\approx$ 60% of the total peptides ($|S^{Train}| = 6760$), S^{Val} contains $\approx$ 20% of the total peptides ($|S^{Val}| = 2253$) and S^{Test} contains remaining $\approx$ 20% of the total peptides ($|S^{Test}| = 2253$). The number of peptides active against each pathogen of the ESKAPEE group out of the total peptides present in D_s is shown in Figure 5.2. Additionally, we have also assured that 60%, 20%, and 20% of the peptides from each pathogen are present in S^{Train}, S^{Val} and S^{Test}, respectively. To determine the efficacy of ABPs against the ESKAPEE group, we obtained the WGS data for the pathogens of the ESKAPEE group from NCBI [108].

5.2.2 Proposed Framework

In the current work, we proposed a framework which utilizes transfer learning and ensemble learning techniques. The concept of ensemble learning was realized by com-

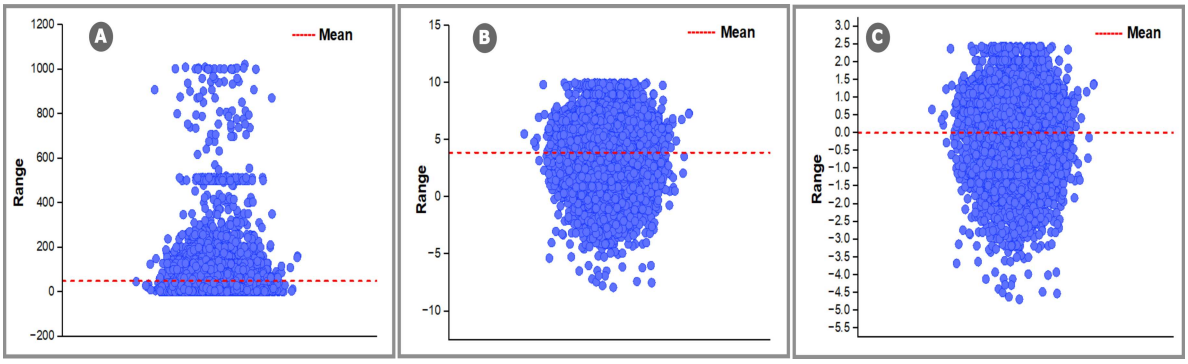

Figure 5.3: Distribution of ABPs (A) across a wide range of MIC values (B) across a narrow range of log (MIC) values (C) across a narrow range of log (MIC) values after standardization.

bining the decisions of deep learning algorithms, namely BiGRU, BiLSTM, BiTCN, and 1DCNN, which have been used in sequence-related tasks in the field of Natural language processing and bioinformatics [109, 76, 110, 77, 111, 78, 112]. The concept of transfer learning was realized by utilizing pretrained amino acid embeddings, namely, PESTV, which provides 1024 dimensional encoded vector corresponding to each amino acid. Our proposed work aims to predict the MIC values against pathogens from the ESKAPEE group for a given peptide. Therefore, we need to input two types of information into our framework: (i) Information about the ABP and (ii) Information about the bacteria against which we want to predict the MIC value. Therefore our proposed framework is a two-input framework that utilizes transfer learning and ensemble learning techniques. The two inputs of proposed framework result in two input branches, namely I_{ABP} and I_{PGN}, which deal with ABP-related information and pathogen-related information, respectively, as shown in Figure 5.1.

In I_{ABP}, the first layer is the Input layer, using which we feed the encoded peptides. The dimensions of the data sent via this layer must match. Since we have considered peptides having a length of $\in$ [5,50], the maximum value for the peptide length is 50. So, for encoded peptides shorter than 50, we performed post-padding using zero vector(s) of length 1024. Therefore, the shape of output from this layer is (b_s, 50,

1024), where b_s stands for batch size (= 16 in our case). Next, corresponding to BiGRU-PESTV, BiLSTM-PESTV, BiTCN-PESTV, and 1DCNN-PESTV models, we added BiGRU, BiLSTM, BiTCN, and 1DCNN layers, respectively. The number of neurons used in each GRU layer of BiGRU is 256, resulting in the output of shape (bs, 50, 512). Next, we used the GlobalMaxPooling1D layer, which downsamples input by taking the maximum value over the time dimension, resulting in the output of shape (b_s, 512). Next, we used the Independent component layer (ICL). The ICL was presented for the first time in [46], where the authors merged batch normalization and dropout. They ran multiple experiments and observed that ICL leads to a stable training process, quicker convergence, and improved generalization. This layer does not change the shape of the input, resulting in the output of shape (b_s, 512). Next, we used two dense layers consisting of 256 neurons and 64 neurons, respectively, resulting in the output of shape (b_s, 256) and (b_s, 64), respectively. In BiLSTM, the number of neurons used with each LSTM layer is 128, resulting in the output of shape (b_s, 50, 256). Next, we used the GlobalMaxPooling1D layer, which performs down-sampling of input by taking the maximum value over the time dimension, resulting in the output of shape (b_s, 256). Next, we used ICL followed by two dense layers consisting of 128 and 32 neurons, resulting in the output of shape (b_s, 128) and (b_s, 32), respectively. In BiTCN, the number of filters used with each TCN layer is 256, resulting in the output of shape (b_s, 50, 512). Next, we used the GlobalMaxPooling1D layer, which performs down-sampling of input, resulting in the output of shape (b_s, 512). Next, we used ICL, followed by two dense layers consisting of 256 neurons and 64 neurons, resulting in the output of shape (b_s, 256) and (b_s, 64), respectively. In 1DCNN, the number of filters used is 256, resulting in the output of shape (b_s, 50, 256). Next, we used the GlobalMaxPooling1D layer, which performs down-sampling of input, resulting in the output of shape (b_s, 256). Next, we used ICL, followed by two dense layers consisting of 128 and 64 neurons, resulting in the output of shape (b_s, 128) and (b_s, 64), respectively.

In I_{PGN}, the first layer is the Input layer, using which we feed the genomic features (GF) of pathogens. For obtaining the GF, we utilized 340 composition-based features (Nucleotide composition (4) + Di-nucleotide composition (16) + Tri-nucleotide composition (64) + Tetra-nucleotide composition (256)) from Nfeature [113], resulting in the output of shape (b_s, 340). Next, we used two dense layers consisting of 128 and 32 neurons, resulting in the output of shape (b_s, 128) and (b_s, 32), respectively.

Next, for each BiGRU-PESTV, BiLSTM-PESTV, BiTCN-PESTV, and 1DCNN-PESTV , we combined the output from corresponding I_{ABP} and I_{PGN} using Concatenate layer, resulting in the output of shape (b_s, 96), (b_s, 64), (b_s, 96), and (b_s, 96), respectively. Next, we used ICL, followed by two dense layers consisting of 32 and four neurons, resulting in the output of shape (b_s, 32) and (b_s, 4), respectively. Next, we used ICL, followed by an output layer consisting of a single neuron, resulting in the output of shape (b_s, 1). Rectified linear unit (ReLU) activation function was used with BiGRU, BiLSTM, BiTCN, 1DCNN, and dense layers

The network weights were updated using the Adam (Adaptive Moment Estimation) optimizer.

We employed the EarlyStopping technique (with parameters patience and epochs as 50 and 5000, respectively) to prevent overfitting. This technique monitors the validation loss and training loss and stops training of the model(s) when the validation loss stops decreasing. Finally, we combined the predictions from the base classifiers BiGRU, BiLSTM, BiTCN, and 1DCNN, which provides us with the actual prediction from our proposed framework, which was further converted to MIC values (in µM) by using the scaler inverse and inverse log transformations.

The variation of training and validation MSE loss for BiGRU-PESTV, BiLSTM-PESTV, BiTCN-PESTV, and 1DCNN-PESTV are shown in Figure 5.4 (A), Figure 5.4 (B), Figure 5.4 (C), and Figure 5.4 (D), respectively. As can be seen from these Figures, both training and validation MSE loss are decreasing with epochs, and the

training of the models is ended when there is no longer a decline in the validation MSE loss. Moreover, the validation and test MSE loss curves are superimposable (as can be seen from Figures 5.4 (A) - 5.4 (D)), which shows that all the models have learned well and there is no overfitting.

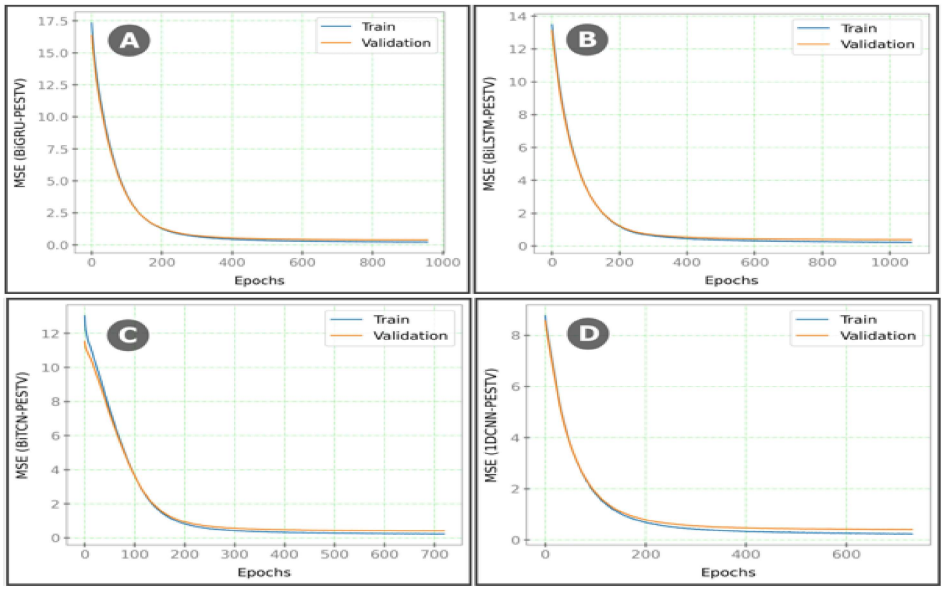

Figure 5.4: The variation of training and validation MSE for BiGRU, BiLSTM, BiTCN, and 1DCNN

5.3 Experiments and Results

This section briefly describes the experimental configuration, performance metrics, assessment procedure used, results obtained from the proposed framework. We have also conducted the ablation studies to understand the contribution of different components in the proposed framework. Moreover, we have also performed additional experiments using HCF. This section also provides the results obtained from these ablation studies and additional experiments.

5.3.1 Experimental Configuration

The deep learning algorithms were implemented using Keras deep learning library [48] with Tensorflow as the backend. All experiments were carried out on a CPU compute node configured with a 2.4 GHz Intel-Xeon Skylake 6148 processor and 192 GB RAM.

5.3.2 Performance Metrics

To access the performance, we used Pearson correlation coefficient (PCC), coefficient of determination (COD), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE), respectively.

5.3.3 Assessment Procedure

Dataset D_s containing 11,266 ABPs was divided into three sets, namely Training set (S^{Train}), Validation set (S^{Val}), and Test set (S^{Test}). S^{Train} contains 60% (6760) peptides . S^{Val} contains 20% (2253) peptides, and S^{Test} contains the remaining 20% (2253) peptides. We used S^{Train} for training, S^{Val} for hyperparameter tuning and identifying the best framework (the model obtained from the best framework is termed as ESKAPEE-MICpred) among the frameworks available from different methods. S^{Test} was retained to test the generalization performance of our proposed model.

Table 5.1: Results obtained from proposed framework.

S.No	Algorithm	Model	PCC	COD	MAE	MSE	RMSE
1	Ensemble	ESKAPEE-MICpred	0.804	0.643	0.446	0.356	0.597

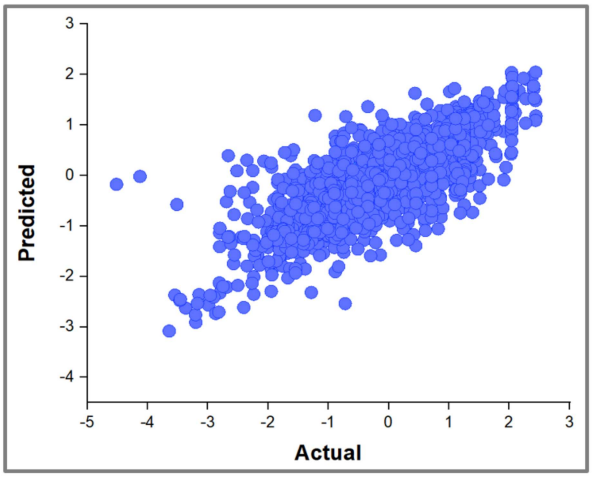

Figure 5.5: Predicted versus actual activity values for peptides.

5.3.4 Results obtained by the proposed Model

We have proposed a model named ESKAPEE-MICpred that utilizes transfer learning and ensemble learning techniques with deep learning algorithms. The results obtained by the ESKAPEE-MICpred for S^{Val} are provided in Table 5.1. This table shows that the value of PCC, COD, MAE, MSE, and RMSE obtained for S^{Val} is 0.804, 0.643, 0.446, 0.356, and 0.597, respectively. Before arriving at the proposed framework, we performed ablation studies and conducted additional experiments . These ablation studies and additional experiments are evaluated on the validation data, and their findings are presented in the subsequent Sections.

5.3.5 Ablation Studies

In the current work, we have utilized (i) weighted average ensemble strategy to combine the predictions of BiGRU-PESTV, BiLSTM-PESTV, BiTCN-PESTV, 1DCNN-PESTV (By experimenting with various weight combinations and assessing their effectiveness on S^{Val}, optimal weights were identified) and (ii) pretrained embeddings from PESTV. Before arriving at the proposed framework, using S^{Train} and S^{Val}, we conducted various experiments with respect to ensemble learning and transfer learning approaches. The

results of these experiments are presented in this Section as ablation studies.

5.3.5.1 Impact of Ensemble learning technique

To understand the contribution of ensemble learning, we eliminated it from the proposed model (ESKAPEE-MICPred). Table 5.2 displays the results obtained for BiGRU-PESTV, BiLSTM-PESTV, BiTCN-PESTV, and 1DCNN-PESTV and Figure 5.6 compares the results obtained by ESKAPEE-MICPred, BiGRU-PESTV, BiLSTM-PESTV, BiTCN-PESTV, and 1DCNN-PESTV. As seen in this Figure, ESKAPEE-MICPred outperformed individual classifiers across all performance metrics. Specifically, PCC and COD values decreased, and MAE, MSE, and RMSE increased when we eliminated ensemble learning.

The models utilised to create the ensemble model are heterogeneous in nature. Therefore, the data point for which one model cannot perform well remaining model can perform well. This might be the probable reason for the better performance of the aggregate model in comparison to individual models.

Table 5.2: Results obtained by BiGRU, BiLSTM, BiTCN and 1DCNN utilising PESTV

S.No.	Algorithm	Model	PCC	COD	MAE	MSE	RMSE
1	BiGRU	BiGRU-PESTV	0.796	0.632	0.453	0.366	0.605
2	BiLSTM	BiLSTM-PESTV	0.791	0.626	0.454	0.373	0.611
3	BiTCN	BiTCN-PESTV	0.783	0.613	0.464	0.386	0.621
4	1DCNN	1DCNN-PESTV	0.787	0.619	0.460	0.380	0.617

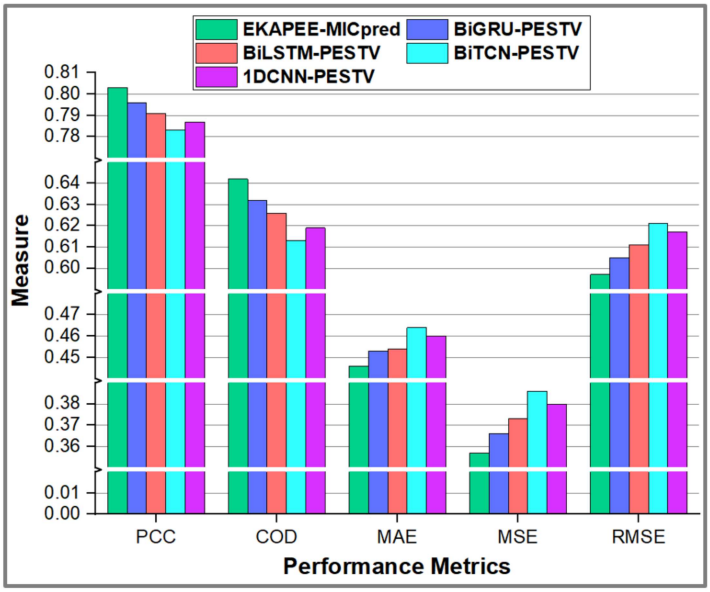

Figure 5.6: Comparison of results obtained by proposed ensemble classifier (ESKAPEE-MICpred), BiGRU-PESTV, BiLSTM-PESTV, BiTCN-PESTV, and 1DCNN-PESTV.

5.3.5.2 Impact of transfer learning

The concept of transfer learning was realised in terms of PESTV. Therefore, to understand the contribution of transfer learning, we replaced PESTV with non-pretrained encodings, namely OHE, PAM 250, BLOSUM 62, and NNAA. NNAA, PAM250, and BLOSUM62 are the most popular amino-acid replacement matrices. In OHE, every amino acid in the vocabulary is represented by a vector containing zeros in all cells except a single cell containing one which uniquely identifies that particular amino acid. By training Embeddings from Language Models (ELMo) on around 33 million protein sequences from UniRef50, authors in [14] learned PESTV. The results obtained from BiGRU, BiLSTM, BiTCN and 1DCNN on utilising OHE, PAM 250, BLOSUM 62, and NNAA are provided in Table 5.3,5.4,5.5,5.6, respectively. The performance obtained by BiGRU, BiLSTM, BiTCN and 1DCNN with different encodings is compared in Figure 5.7, 5.8, 5.9, and 5.10, respectively. As can be seen from these figures, the performance of all the classifiers degraded when we replaced PESTV with other non-pretrained en-

codings.

The probable reason for the better performance of PESTV compared to non-pre-trained embeddings is transfer learning, where the idea is to leverage the knowledge gained while solving one problem and apply it to a different but related problem. In our case, the knowledge acquired by PESTV on 33 million protein sequences is utilised in predicting the MIC value for ABP.

Table 5.3: Results obtained by BiGRU, BiLSTM, BiTCN and 1DCNN utilising OHE

S.No.	Algorithm	Model	PCC	COD	MAE	MSE	RMSE
1	BiGRU	BiGRU-OHE	0.747	0.557	0.509	0.441	0.664
2	BiLSTM	BiLSTM-OHE	0.759	0.574	0.496	0.424	0.651
3	BiTCN	BiTCN-OHE	0.774	0.599	0.471	0.400	0.632
4	1DCNN	1DCNN-OHE	0.752	0.561	0.506	0.437	0.661

Table 5.4: Results obtained by BiGRU, BiLSTM, BiTCN and 1DCNN utilising PAM250

S.No.	Algorithm	Model	PCC	COD	MAE	MSE	RMSE
1	BiGRU	BiGRU-PAM	0.769	0.590	0.483	0.409	0.639
2	BiLSTM	BiLSTM-PAM	0.767	0.588	0.482	0.411	0.641
3	BiTCN	BiTCN-PAM	0.782	0.610	0.467	0.389	0.623
4	1DCNN	1DCNN-PAM	0.757	0.567	0.506	0.431	0.657

Table 5.5: Results obtained by BiGRU, BiLSTM, BiTCN and 1DCNN utilising BLOSUM62

S.No.	Algorithm	Model	PCC	COD	MAE	MSE	RMSE
1	BiGRU	BiGRU-BLOSUM	0.772	0.596	0.484	0.403	0.634
2	BiLSTM	BiLSTM-BLOSUM	0.776	0.602	0.467	0.396	0.630
3	BiTCN	BiTCN-BLOSUM	0.778	0.605	0.472	0.394	0.628
4	1DCNN	1DCNN-BLOSUM	0.758	0.569	0.506	0.430	0.656

Table 5.6: Results obtained by BiGRU, BiLSTM, BiTCN and 1DCNN utilising NNAA

S.No.	Algorithm	Model	PCC	COD	MAE	MSE	RMSE
1	BiGRU	BiGRU-NNAA	0.762	0.580	0.492	0.419	0.647
2	BiLSTM	BiLSTM-NNAA	0.774	0.598	0.479	0.401	0.633
3	BiTCN	BiTCN-NNAA	0.779	0.605	0.473	0.393	0.627
4	1DCNN	1DCNN-NNAA	0.724	0.512	0.545	0.487	0.697

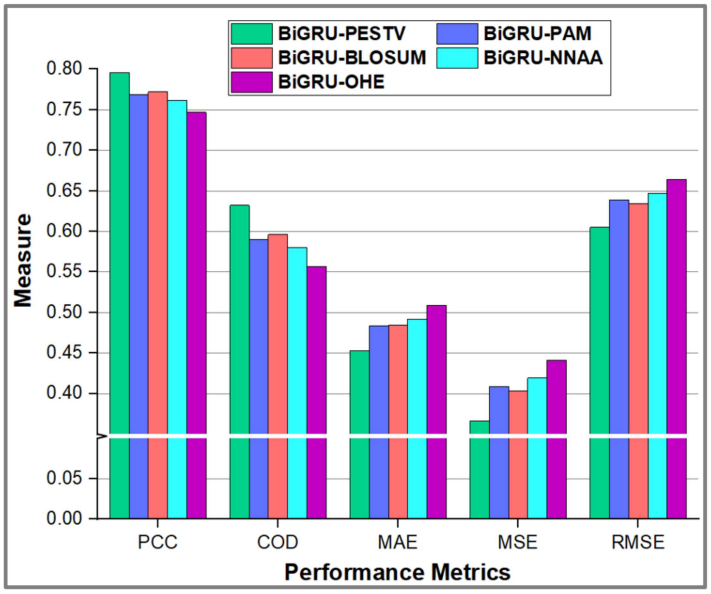

Figure 5.7: Comparison of results obtained by BIGRU algorithm when it is utilized with pretrained embeddings (PESTV) and non-pretrained embeddings (PAM 250, BLOSUM 62, NNAA, and OHE).

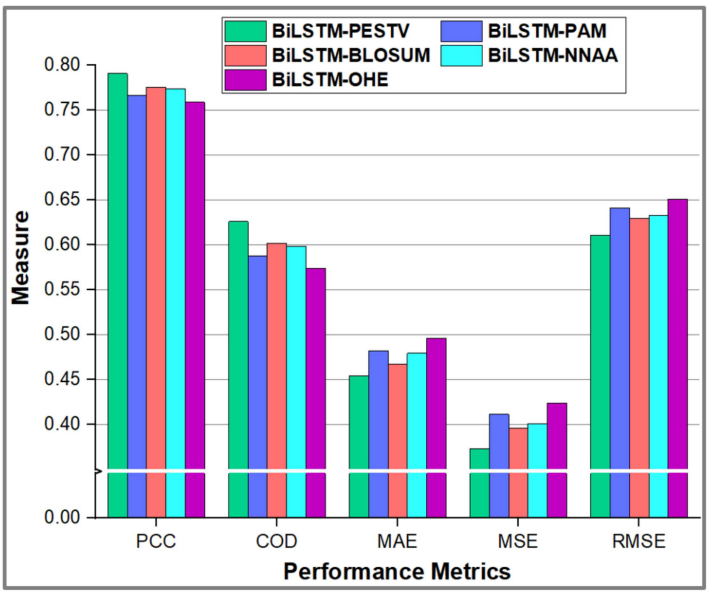

Figure 5.8: Comparison of results obtained by BILSTM algorithm when it is utilized with pretrained embeddings (PESTV) and non-pretrained embeddings (PAM 250, BLOSUM 62, NNAA, and OHE).

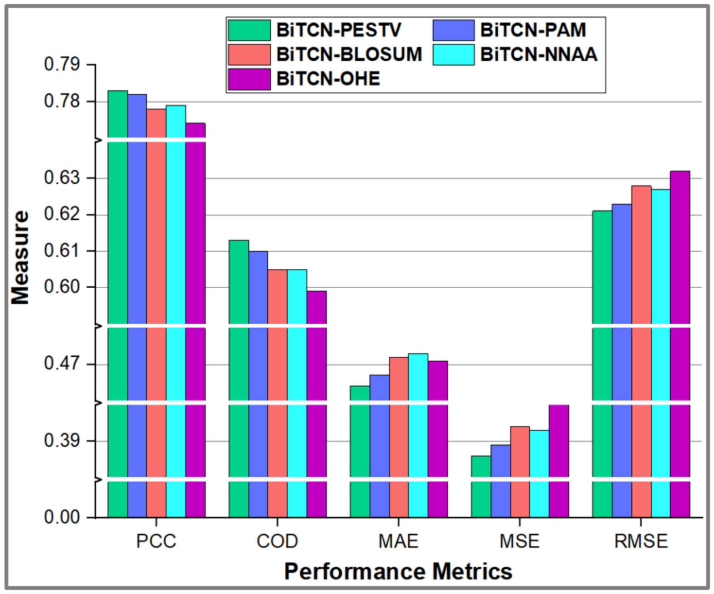

Figure 5.9: Comparison of results obtained by BITCN algorithm when it is utilized with pretrained embeddings (PESTV) and non-pretrained embeddings (PAM 250, BLOSUM 62, NNAA, and OHE).

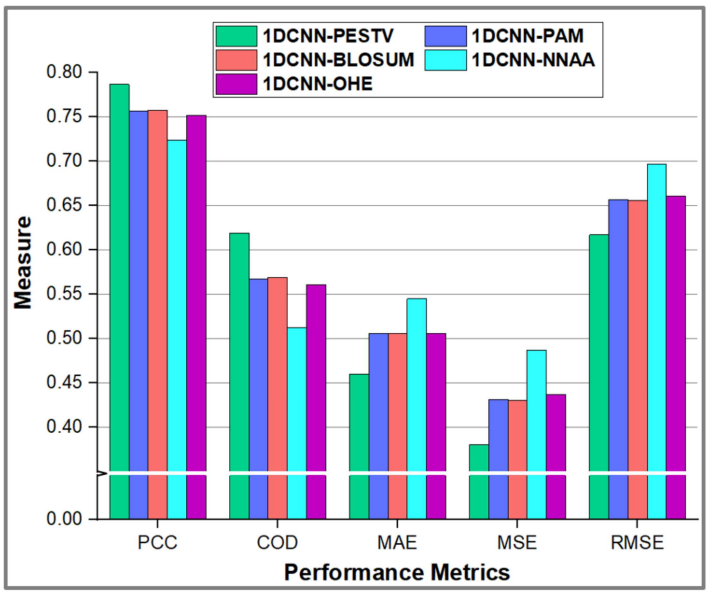

Figure 5.10: Comparison of results obtained by 1DCNN algorithm when it is utilized with pretrained embeddings (PESTV) and non-pretrained embeddings (PAM 250, BLOSUM 62, NNAA, and OHE).

5.3.6 Additional Experiments

Additional experiments include utilising the handcrafted features (HCF). For preparing HCF, we used the protr package available in R, which provided 9821 features corresponding to each peptide, which were further reduced to 271 features by removing constant features, duplicate features and correlated features. We have utilised these HCF in two ways : (i) With Artificial Neural Network (ANN) (ii) With BiGRU, BiLSTM, BiTCN and 1DCNN in combination with PESTV.

5.3.6.1 Utilising HCF with ANN

We have created a feature vector of length 611 by combining 271 HCF with 340 genome-level features (which provide information about the ESKAPEE bacteria). This feature vector was further utilised with ANN. The results obtained from here are provided in Table 5.7. As can be seen from this Table, the results obtained by utilising HCF with ANN are inferior to the results obtained by our proposed model ESKAPEE-MICpred

(S.No.1 Table 5.1). This shows that the DLF used in our proposed model contains more information in comparison to HCF, which makes our proposed model perform better than ANN-HCF.

Table 5.7: Results obtained by ANN utilising HCF

S.No.	Algorithm	Model	PCC	COD	MAE	MSE	RMSE
1	ANN	ANN-HCF	0.751	0.559	0.506	0.439	0.663

5.3.6.2 Utilising both HCF and PESTV with BiGRU, BiLSTM, BiTCN and 1DCNN

Sometimes HCF includes additional information that is not there with the DLF and vice versa; in such cases, it is possible to improve the performance by combining both HCF and DLF. Therefore, we have also experimented by combing the HCF with DLF. The results obtained are provided in Table 5.7. But we did not observe any improvement by doing so. This shows that HCF did not add any additional information, and PESTV alone are capable of doing the desired task.

Table 5.8: Results obtained by BiGRU, BiLSTM, BiTCN and 1DCNN utilising BOTH HCF AND PESTV

S.No.	Algorithm	Model	PCC	COD	MAE	MSE	RMSE
1	BiGRU	BiGRU-PESHCF	0.784	0.615	0.460	0.383	0.619
2	BiLSTM	BiLSTM-PESHCF	0.786	0.617	0.458	0.382	0.618
3	BiTCN	BiTCN-PESHCF	0.777	0.601	0.476	0.397	0.630
4	1DCNN	1DCNN-PESHCF	0.777	0.598	0.481	0.400	0.633

Table 5.9: Results obtained from the proposed model ESKAPEE-MICpred on test data.

S.No	Algorithm	Model	PCC	COD	MAE	MSE	RMSE
1	Ensemble	ESKAPEE-MICpred	0.802	0.640	0.442	0.376	0.613

Table 5.10: Antibacterial peptides identified from the proteins sequences.

Protein ID	Organism	Sequence	Length	MW	MIC against bacteria(in µM)						
					E	S	K	A	P	E	E
Q75VN4	Zhangixalus schlegelii	AKKGSKKAVSKVQKKDGKKRRKSRK	25	2855.44	19.86	20.02	27.43	20.87	19.40	18.79	10.97
F8J4S0	Oxyopes takobius	RCPKSWKCKAFKQRVLKRLLAMLR	24	2960.73	5.55	7.03	7.96	4.84	9.23	6.97	5.70
KAF7238795	Varanus komodoensis	PATGGVKKPHRYRPGTVALREIRRY	25	2879.33	10.00	14.23	19.32	13.42	18.18	17.23	13.63
XP044279191	Varanus komodoensis	KKGSKKAITKTQKKGNKKRKKSRK	24	2785.39	12.10	17.45	11.63	8.31	9.37	7.67	4.32
KAF7243928	Varanus komodoensis	LTVHLRKKHQFKWPSGHPRFRYK	23	2947.45	5.09	6.30	13.57	6.91	12.70	11.00	7.13

Table 5.11: Antibacterial peptides identified from the therapeutic peptides.

DBAASP ID	Existing property	Sequence	Length	MW	MIC against bacteria(in µM)						
					E	S	K	A	P	E	E
12592	Anticancer	FKKLKKLFSKLWNWKRKKRRQRRR	24	3316.06	3.99	4.58	7.45	4.64	6.51	6.50	4.50
3052	Anticancer, Antifungal	RIIDLLWRVRRPWKPKFVTVWVR	23	3020.67	2.07	0.59	8.15	3.91	5.52	6.16	4.35
4944	Antifungal	RWRSFFKKAAHRGKHVGKRARTHYL	25	3094.59	3.55	4.20	5.09	4.89	3.56	5.48	3.30
3053	Anticancer, Antifungal	RIIDLLWRVRRPQWPKFVTVWVR	23	3020.63	1.41	0.46	9.23	3.44	5.65	5.83	3.76
13339	Antiviral	FLFAFRIFKRVFKKFRKLFKRAF	23	3041.77	2.40	2.86	6.30	2.18	6.60	3.72	2.69

5.3.7 Performance of the proposed model ESKAPEE-MICpred on test data

By performing ablation studies and additional experiments, we found that our proposed model ESKAPEE-MICpred is better than others. Further, we evaluated the generalization performance of our proposed model on S^{Test}. The results obtained by the ESKAPEE-MICpred for S^{Test} are provided in Table 5.9, which shows that the value of PCC, COD, MAE, MSE, and RMSE obtained for S^{Test} is 0.802, 0.640, 0.442, 0.376, and 0.613, respectively. We can see that the values of all the performance metrics in the case of S^{Test} is same as that of S^{Val} , which depicts the stability of the proposed model. We have also provided the plot between actual and predicted values for S^{Test} in Figure 5.5. As can be seen from this Figure, there are few data points in the second and fourth quadrants of the graph, which shows that there are few peptides that are predicted to be highly active but are actually not and vice versa.

5.4 Prediction of MIC values against ESKAPEE

Our proposed model can be used to predict ABP's MIC values against the ESKAPEE group. As a pilot study, we considered five antibacterial protein sequences for identifying ABPs and proposed one ABP from each of these five proteins, which can target all the ESKAPEE pathogens at low MIC (details are provided in Table 5.10). For this, we performed the following steps: (i) For each of the five proteins, we created the peptide library by obtaining the substrings of length $\in$ [20, 25]. (ii) We fed the peptide library into ABP classifiers, namely Deep-ABPpred [75], AMPFUN [114], MSTCN-ABPpred [115], AMPScanner [116] (iii) The peptides which are classified as ABPs by all the models are utilized with BiGRU-PESTV, BiLSTM-PESTV, BiTCN-PESTV, and 1DCNN-PESTV (iv) The results obtained from (iii) are ensembled together, and then scaler inverse and inverse log transformations are taken. Finally, we proposed five

peptides (one from each protein) that are soluble in water and are non-hemolytic (Solubility is determined using http://pepcalc.com/, and hemolytic activity is determined using [4]) for wet lab synthesis and experimentation.

In addition, our proposed model can be utilized to examine therapeutic peptides that have not been experimentally evaluated for antibacterial activity, which will help in drug repurposing. As a pilot study, we considered such therapeutic peptides from DBAASP and proposed five peptides for wet lab synthesis and experimentation utilizing steps (ii) to (iv). (details are provided in Table 5.11)

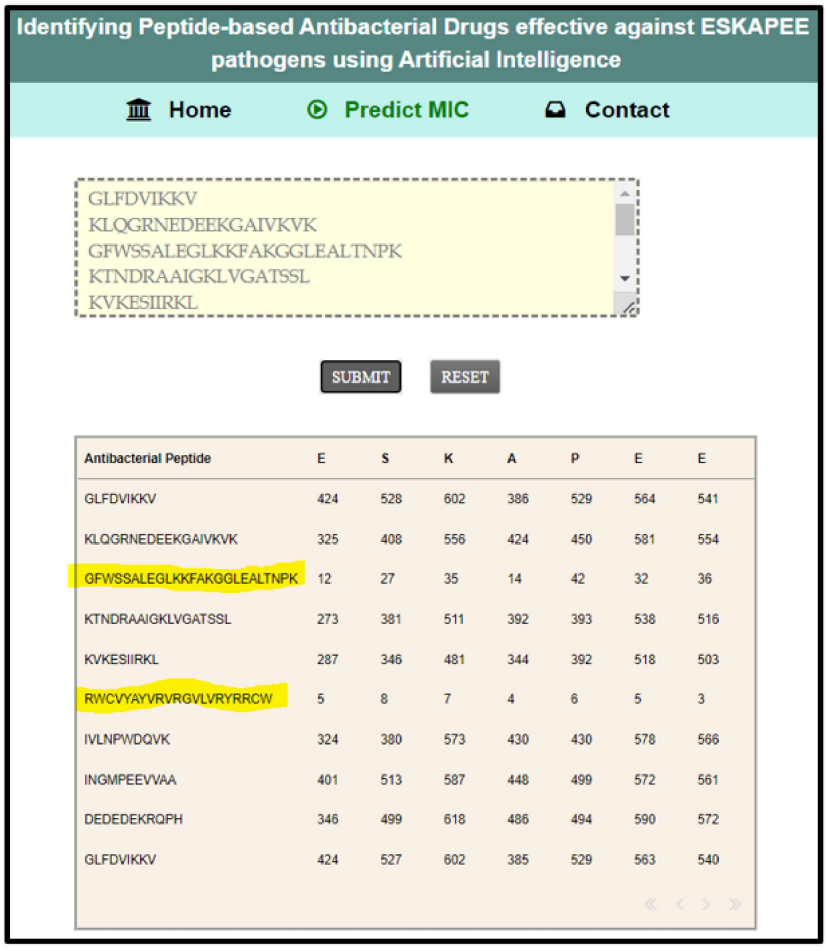

Figure 5.11: Predicting the MIC value of antibacterial peptides against ESKAPEE pathogens.

5.5 Web Server

To serve the scientific community, we have made ESKAPEE-MICpred accessible online in the form of a web server at https://eskapee-micpred.anvil.app/. The web server accepts as input ABP of length $\in$ [5, 50] and returns the MIC values of ABP against the EKKAPEE group of pathogens. The MIC values obtained for some of the sample ABPs are shown in Figure 5.11

5.6 Summary

The ESKAPEE group of bacteria, which is the main source of AMR, has drawn a lot of attention from the WHO. ABPs are a family of peptides found in nature that play a crucial role in the innate immune systems of organisms. These ABPs have a broad spectrum of activity; as a result, bacteria cannot develop resistance against them. Therefore, ABPs have recently received much attention as a potential replacement for currently available antibiotics. But it is expensive and time-consuming to identify ABPs from natural sources. Thus, the wet lab researchers employ various computational tools available in the public domain to rapidly screen promising ABPs. However, the main limitation of the existing tools is that they do not provide the MIC values for the identified ABP against the ESKAPEE pathogens. Therefore, after identifying ABPs, wet lab researchers have to test them against the ESKAPEE pathogens at different concentrations, leading to a loss of time and money. Moreover, it is only possible to test some of the identified ABPs in the lab at different concentrations due to the involvement of time and money. This lead to the loss of optimal ABP(s), which can work at low MIC(s) against all the ESKAPEE group of bacteria. To address this, in the current work, we developed ESKAPEE-MICpred, which provides MIC values for the ABPs against the ESKAPEE group. For developing ESKAPEE-MICpred, we utilised the concept of transfer learning and ensemble technique. The concept of trans-

fer learning was realised with the help of PESTV, whereas the ensemble technique was applied by combining the predictions made by BiGRU, BiLSTM, BiTCN and 1DCNN deep learning algorithms. We have obtained a PCC value of ≈ 0.8, which shows a good correlation between the actual and predicted MIC values. We have also performed ablation studies to justify the usage of transfer learning and ensemble learning techniques. As a pilot study, we considered histone proteins from the Komodo dragon and suggested five peptides (one from each category of histone protein) which are effective against the ESKAPEE group. The proposed model has been deployed as a web server at https://eskapee-micpred.anvil.app/ to aid the scientific community. We hope this tool will aid wet lab researchers in discovering novel potent ABPs against ESKAPEE pathogens and help combat the alarming situation of AMR.

There is room to improve the model by incorporating the data-centric approach, which focuses on enhancing the dataset in terms of both quantity and quality. ABPs have recently garnered considerable interest as a potential alternative to currently available antibiotics. As a result, experimentally validated ABPs are increasing at a very fast rate. Therefore we can improve the quantity of data by incorporating such newly discovered data. As we have already mentioned in Section 5.2.1 multiple studies have reported different MIC values for the same ABP due to varying experimental conditions. Therefore we can also improve the quality of data by involving a domain expert who can review each data sample and eliminate any out-of-place data points.